The National Center for Biomedical Ontology

Enabling the retrieval, integration, and analysis of "big data" in biomedicine

In June 2012, the Data and Informatics Working Group reporting to the Advisory Committee to the Director recommended that the NIH "promote data sharing through central and federated catalogs." The foundation of such catalogs for data sharing rests in the ontologies and controlled terminologies that provide the standard definitions of the elements of biomedical data sets and of the contents of electronic health records. Without standard mechanisms to define the meaning of data, it is impossible to retrieve, integrate, and analyze the vast data sets that are now the norm in biomedicine and that form the basis of electronic health records.

The National Center for Biomedical Ontology (NCBO) has become the leading scientific organization for bringing this sort of semantic technology to biomedicine. The Center has its main activity at Stanford University (PI: M.A. Musen), with collaborators at the Mayo Clinic (co-PI: C.G. Chute), the University at Buffalo (co-Pi: B. Smith), and the University of Victoria (co-PI: M.-A. Storey).

Technology: The NCBO has created the world's definitive online ontology repository, known as **BioPortal**, which stores some **350 biomedical ontologies and controlled terminologies** in a uniform, consistent manner. NCBO makes these disparate ontologies accessible in a standardized manner through a Web browser interface that attracts more than **65,000 visitors per month**—thousands of whom visit the site every working day.

The Center provides programmatic access to its ontology repository through Web services that are invoked by application programs in the laboratory and in the clinic. The programmatic interface to BioPortal receives a median of **1 million calls per month**. The NCBO offers additional Web services that put the BioPortal repository to use for recurring biomedical tasks that make use of our ontology content. The **NCBO Annotator**, for instance, is a high-throughput Web service that takes as input some text (such as a PubMed abstract or the textual metadata of a data set in an online repository such as GEO) and returns as output the terms from preselected ontologies that relate to that text. Thus, the ArrayExpress resource at the European Bioinformatics Institute makes extensive use of the NCBO Annotator to link the metadata associated with microarray studies to standard ontologies in BioPortal, enabling high-precision indexing, searching, and integration of gene-expression data for many thousands of users. At Stanford, investigators have applied the NCBO Annotator to electronic health records, analyzing the text of thousands of progress notes and discharge summaries *per minute* to identify adverse drug events and off-label drug usage. In a test to search for signals involving 9 drugs that caused known adverse events, the NCBO Annotator enabled identification of 7 of these—on average 1.9 years before the FDA called for the drugs to be withdrawn from the market.

User Community: The NCBO serves a vast and extremely vibrant user community that has developed more than **55 third-party biomedical applications** that rely on NCBO Web services for their infrastructure; our users created 15 of these applications in the last year alone. For example, the WHO relies on NCBO technology in its tools to develop the new revision of the International Classification of Diseases, aligning the terms in ICD-11 with those in SNOMED CT. The i2b2 data warehouse uses NCBO Web services to populate its "ontology hive." Investigators at UCSD collaborated with Microsoft Research to create a plug-in for Microsoft Word that calls NCBO Web services as authors are working on a document to mark-up the text with appropriate ontology terms (for example, linking the name of an enzyme to a specific entry in the Protein Data Bank). caNanoLab is a resource that makes extensive use of NCBO technology to allow the NCI's nanotechnology community to share precise information about the physical, chemical, and biological characteristics of therapeutic nanoparticles.

Training and Dissemination: The NCBO has a very active training and dissemination effort. We have attracted **16 visiting scholars,** and we have trained **18 graduate students** and **16 post-doctoral fellows** in the use of semantic technology in biomedicine. Our trainees have gone on to promote the use of NCBO technology in academia and in industry. We have organized **59 educational workshops and conferences** and have presented **17 tutorials** at national and international meetings. We regularly host webinars on the use of our technology. There have been **25 webinars** the past 2 years; some **3,500 people** have accessed our online recordings of these talks.

The National Alliance for Medical Image Computing

Quantitative image analysis is essential for extracting knowledge from biomedical images. The National Alliance for Medical Image Computing (NA-MIC) is a multi-institutional NCBC that investigates algorithms and develops tools vital for translational research in a modular, open-source software infrastructure.

RESEARCH IMPACT: *NA-MIC has published 456 papers in peer review journals and conference reports, including several award-winning papers presented at high-end conferences.* The methodology developed by NA-MIC scientists is driven by 11 Driving Biological Projects that address neurodegenerative disorders (schizophrenia and autism), lupus, Huntington's disease, heart disease (atrial fibrillation), radiotherapy for prostate and head and neck cancers, and traumatic brain injury. In addition to scholarly papers, the NA-MIC community makes this software available in the NA-MIC Kit and in 3D Slicer, an established software application used worldwide that enables research in the engineering sciences and in biomedical research.

Mobilizing an International, Open-Source, Development Community. *NA-MIC has adopted an open and inclusive approach to building a community of scientists around the NA-MIC Kit.* In addition to a variety of US activities with NIH grantees, contributions to the NA-MIC Kit have come from substantial government-funded efforts in Canada, Germany, Spain, France, and Italy. Examples of NA-MIC's impact on the international community include the Ontario Consortium for Adaptive Interventions in Radiation Oncology (OCAIRO), which was awarded Canadian funding in 2011 to develop research software for adaptive radiation therapy based on the NA-MIC Kit. The results of this work are made available to the NA-MIC community as extensions. In a similar vein, NA-MIC investigators have co-led the creation of the Common Toolkit (CTK) effort, an international collaboration with substantial contributions from several European countries.

Combining the Analysis of Genetic and Imaging Data. *NA-MIC's work with the NIH-funded COPDGene project (www.copdgene.org) exemplifies the center's impact on collaborative research.* NA-MIC technology has enabled the correlation of imaging-derived quantitative measures of airway, parenchymal, and vascular phenotypes with a spectrum of established pulmonology diagnostic metrics and a genome-wide association study. This analysis has established that genes near *CHRNA*-3/5 and *MMP*-12/13 can determine the genetic predisposition of an individual to develop COPD. This same conceptual approach of combining quantitative image analysis based on the NA-MIC Kit with genetics is being used in a second multi-center study called PREDICT-HD which is run by a Huntington's disease (HD) consortium led by the University of Iowa. The PREDICT-HD consortium uses the NA-MIC Kit to research quantitative medical imaging bio-markers as surrogate endpoints in drug treatment trials that are aimed at delaying disease onset and/or progression.

COMMUNITY RESOURCES/SOFTWARE/COLLABORATIONS: *NA-MIC's algorithms and tools have been broadly adopted by the community and by industry.* 3D Slicer was downloaded 41,000 times in the past 12 months. 3D Slicer's user and developer mailing lists contain 829 and 483 members, respectively. CMake, a multi-platform developer environment, is one of the most popular components of the NA-MIC Kit, and has an impact beyond the biomedical field. This package has more than 2,000 known downloads/day. NA-MIC has 31 funded collaborations: these include 25 NIH grants (8 active, 17 completed) and 6 international grants (5 active, 1 completed). These collaborations address a broad range of organ systems and pathologies: diagnosis and therapy of schizophrenia, lupus erythematodes, autism, lung disease, cardiac disease, cancer of the brain, liver, colon, and prostate, and musculoskeletal disorders.

TRAINING & DISSEMINATION: *NA-MIC scientists have directly mentored over 55 software engineers, 35 doctoral students, and 20 postdoctoral fellows.* In addition, NA-MIC has trained over 2,000 investigators in the use of 3D Slicer and other components of the NA-MIC Kit through 63 hands-on workshops. To complement these customized events, NA-MIC has developed a freely available online training compendium, consisting of 88 detailed tutorials. A different type of research is presented in the grand challenge workshops at premier conferences, such as the pioneering initiative on standardized evaluation of diffusion tensor imaging tractography algorithms for neurosurgical planning at MICCAI. Finally, NA-MIC practices the best principles of collaborative science through its semi-annual Project Week events. To date, it has held 15 consecutive week-long events, where experts and students gather to address current research problems. Each of these events attracts more than a hundred participants, many of whom return year after year. The hands-on format is extremely popular and has been recognized and adopted by several other centers.

Informatics for Integrating Biology and the Bedside (i2b2): A Translational Engine at the National Scale

RESEARCH IMPACT: Since its inception in 2004 i2b2 has been designed to provide the instrumentation for ***using the informational byproducts of health care and the biological materials accumulated through the delivery of health care to*** – and as a complement to prospective cohort studies and trials - ***conduct discovery research and to study the healthcare system in vivo***. The utility of this approach is demonstrated by the grass-roots adoption of the i2b2 platform by over 84 academic health centers (AHCs) internationally, each implementation of which represents a major, local institutional commitment. **IMPACT EXAMPLE 1. Genomic Disease Studies**. As presented in our recent Nature Genetics Review, the field of Electronic Health Record (EHR) Driven Genomic Research (EDGR) has come into its own. We have made significant contributions with our validations of findings made in other studies in a broad array of phenotypes (e.g. RA, MDD, Asthma, IBD). In all studies the directionality of the odds ratios of SNPs reproduced with magnitudes within 95% confidence limits, all at least 1-2 orders of magnitude *faster and cheaper*. We were furthermore able to measure the effect size of SNPs in minority populations due to overrepresentation in AHC EHRs. **IMPACT EXAMPLE 2. Pharmacovigilance (as a public health application).** Our team has successively used EHR mining to confirm the association of increased MI mortality with Vioxx use, the elevated MI risk with Avandia usage (contributing to the FDA "black box" warning), and in collaboration with SIMBIOS and Vanderbilt to rapidly confirm an FDA alert regarding increases in blood glucose in patients taking both paroxetine and pravastatin. **Publications.** The core i2b2 team has produced 185 peer reviewed papers exclusive of 121 publications directly resulting from our Natural Language Processing (NLP) Challenges and including over a dozen in journals with impact factors of 20 and higher. Several of them are the first of their kind in demonstrating direct utility of EHR data.

COMMUNITY RESOURCES/SOFTWARE/COLLABORATIONS: i2b2's software platform ("i2b2"), designed to enable discovery research from existing healthcare information, has provided a major leveraging factor for population-based studies, quality care and outcomes initiatives, adverse event monitoring, and novel hypothesis-driven investigations. The freely available i2b2 toolbox is now used and is being extended by an Academic Users' Group (AUG) now numbering over 300 members. Attendance at our Annual Users' Group Conference, which now exceeds 125 members representing all key constituencies (CTSAs, AHCs, HMOs, Industry, Disease Networks), affirms the value of this product and the community collaboration that has developed to push it forward. i2b2 has led with the Harvard CTSA the development and deployment of a web-based network, SHRINE, that enables data sharing across i2b2 (or other) platforms, as exemplified by the University of California's recent deployment of SHRINE to allow the analysis of 11 million patients across their AHC's. To advance the essential NLP tools necessary to crisply define phenotypes derived from clinical data, we have developed and hosted 6 International NLP Challenges based on annotated, de-id'd patient data sets that have resulted in participation by 138 international teams from 11-45 organizations, a Research Data Set available from our website that has 3,117 unique downloads from a user base of 230 academic researchers, graduate students, industry, and course developers, and 121 publications.

TRAINING & DISSEMINATION: i2b2's Ed/Dissemination Core has prioritized the recruitment of talented undergraduate students into graduate study in the area of healthcare informatics by establishing a Summer Institute in Bioinformatics and Integrative Genomics that has graduated 94 students (8 programs), including 35 URMs. Of the 64 who have graduated college, 42 are now in MD, PhD or MD/PhD programs, including 16 URMs. We have in addition participated in the training of 30 graduate students and 33 postdocs with a stable core faculty size of 15. 20 of the postdocs now hold faculty appointments, 10 are still in training. Support to our user communities is provided by an active Community Wiki, AUG listserv, twice yearly software workshops, and annual AUG Conferences, NLP Workshops, NLP Challenges, and SHRINE National Conferences. www.i2b2.org.

MAGNet – National Center for the Multiscale Analysis of Genomic & Cellular Networks

The Center for The Multiscale Analysis of Genomic & Cellular Networks (MAGNet) was established in 2005, with the mission of providing the research community with novel, Structural and Systems Biology methods and tools for the ***dissection of molecular interactions in the cell*** and for the ***interaction-based elucidation of cellular phenotypes***. A key component of this mission was the validation of these tools through collaborative projects with experimental biologists, whose scientific goals could not have been accomplished without them. These goals were largely exceeded and MAGNet has developed into a major center in Computational Structural and Systems Biology, producing both high-impact science and valuable software tools for the research community. Objective criteria supporting this statement include: number and quality of scientific publications, funded collaborations supported by MAGNet tools, software downloads, utilization criteria, and impact on the Systems Biology community via MAGNet organized activities. MAGNet has also had a profound impact on the academic environment at Columbia University, providing the impetus for the creation of *a new Department of Systems Biology*, under the leadership of Drs. Califano and Honig. Briefly, some key accomplishments of our Center include:

RESEARCH IMPACT: Since 2005, *MAGNet supported research produced 261 papers, including 86 in journals with Impact Factor (IF) ≥ 9.38 (PNAS). Of these, 27 were published in journals with IF ≥ 22.97 (Nat. Biotech.).* The results presented in these publications were often of a seminal nature, including for instance (a) the elucidation of the role of DNA shape in protein-DNA binding specificity, the identification of the Master Regulators of the mesenchymal subtype of Glioblastoma, and the discovery of an extensive microRNA-mediated regulatory network of RNA-RNA interactions in brain tumors. These findings were driven by the Center's Driving Biological Projects (DBPs, http://magnet.c2b2.columbia.edu/?q=node/6) and were the result of close collaboration between experimental and computational biologists. In all cases, the computational methods and software tools developed by MAGNet investigators were instrumental in enabling the scientific discovery.

COMMUNITY RESOURCES/SOFTWARE/COLLABORATIONS: MAGNet algorithms and tools have been broadly adopted by the community. We have designed new (often first-of-a-kind) methods for the dissection of transcriptional, post-translational, genotype-phenotype, and cell-cell interactions, as well as for regulatory-network based analysis of cell phenotypes. In addition to being independently available from the originating investigator labs, these algorithms have been implemented in the *geWorkbench* platform, which has been *downloaded more than 10000 times by more than 800 unique user*s. Attesting to its community impact, *geWorkbench* has been supported by one of NCI's Knowledge Centers, independent of MAGNet support, allowing comments, bug fixes, and new functionality requests to be continuously tracked via user discussion groups. Individual tools, such as ARACNe, have also been downloaded thousands of times and have been incorporated in other platforms, such as the Minet Bioconductor package and GenePattern. On aggregate, the top 5 most popular Center tools *have been downloaded or visited (for web-based services) more than 38,000 times*. Furthermore, *MAGNet methods and tools have been used in numerous biological projects, including the 10 Center DBPs and 73 collaborative projects, of which 39 have resulted in NIH funded activities*. These have been instrumental in validating the computational methods and in demonstrating their value to address important biomedical problems, especially in oncology.

TRAINING & DISSEMINATION: MAGNet Education and Dissemination Cores have achieved significant impact on the training of computationally-savvy structural and systems biologists and in fostering dialogue at the interface between computational and experimental sciences. *Over 100 pre-doctoral students and post-doctoral fellows currently work in MAGNet investigator labs* (61 have received MAGNet funding), and they benefit from the interdisciplinary environment that has been created. Additionally, *we have organized and developed important conferences and meeting*s that attract hundreds of scientists each year, including (*a*) the DREAM conference, to establish objective, community-based benchmarks to test reverse-engineering algorithms, (*b*) the RECOMB Systems Biology conference, which brings together a community of close to 500 researchers, (*c*) the NY Academy of Science Systems Biology Interest Group, and (*d*) the Keystone meeting on Biomolecular Interaction Networks: Function and Disease. Finally, MAGNet provides key informatics support to Columbia's Clinical and Translational Science Award program and to the Herbert Irving Comprehensive Cancer Center. As a result of the impact that MAGNet has had to biomedical research both at Columbia University and at the national level, *in 2010 the University approved the creation of a new department of Systems Biology* to consolidate and streamline the Center's research and education activities. This has enabled us to hire several talented faculty members (Dr. Saeed Tavazoie from Princeton, Drs. Sagi Shapira and Peter Sims from Harvard, Dr. Yufeng Shen from Columbia, and Dr. Chaolin Zhang from Rockefeller) whose arrival is further increasing the capacity of our center and extending our research interest to exciting new fields, including single cell studies, neurobiology, and infectious diseases.

iDASH

integrating Data for Analysis, Anonymization and Sharing
A National Center for Biomedical Computing
http://idash.ucsd.edu

iDASH is the newest National Center for Biomedical Computing, funded in late 2010. Its goal is to develop infrastructure, services, and tools to allow privacy-preserving data sharing. The Working Group on Data and Informatics has recently made recommendations to the NIH Advisory Committee to the NIH Director to accelerate research: Data and meta-data should be shared, incentives should be offered to those who share data, and investments in user training and infrastructure need to be coordinated to ensure efficient utilization of resources. On the training side, the number of informatics professionals and researchers needs to increase. On the infrastructure side, a backbone for data and software sharing needs to be implemented through a network of biomedical computing centers. iDASH addresses both challenges. We are exploring how biomedical researchers and healthcare providers can remain focused on their activities and outsource data storage, de-identification, annotation, curation, some analysis, and distribution to reliable third parties/processes.

Research Impact

iDASH has developed different models, tools and infrastructure for data sharing that allows it to broker the relationship between data owners and data users. The infrastructure, service and tools developed by iDASH protect the privacy of individuals and of institutions, and provide meaningful information for patients to make informed decisions about sharing their data and specimens. We have developed a HIPAA-compliant hardware and software infrastructure at the San Diego Supercomputer Center that combines over 300 terabytes of cloud storage with high performance computing to allow computation on sensitive data such as human genomes and clinical records. Our infrastructure is supporting advanced research in cloud computing and privacy technology that involves commercial and private HIPAA-compliant clouds.

EXAMPLE 1: We have deployed cloud storage computing and associated policy infrastructure for researchers to share data. We are hosting several data sets including different data modalities (whole genomes, transcriptome data, images, specialty reports, clinical trial data, structured and unstructured clinical data) in our annotated data repository, including many related to Kawasaki Disease, a relatively rare disease of unknown etiology for which we have one of the world's largest data collections, which is annotated and mapped to public ontologies using tools from NCBO and other tools that we have developed. **EXAMPLE 2:** Our data sharing models also include facilitating access to federated databases. We host the hub for five University of California health systems, a collection of 11 million patients. Our tools complement our implementation of i2B2 software for count queries with analytical software for privacy-preserving predictive model building. We have also enabled policy-based data exchanges by developing a legal framework of data-use agreements (DUA) between both (a) data providers and iDASH as data custodian (i.e., honest broker similar to an escrow service), and (b) data recipients and iDASH. These DUAs allow the provider to precisely specify what is shared and when (e.g., embargo prior to article publication), the sensitivity of the data (e.g., identified vs. de-identified), and restrictions on who can access the data with a fine control. We have executed over 15 DUAs and this number is increasing fast since our deployment in March 2012. We also developed an electronic informed consent tool to allow patients to express their preferences towards the use of their data and institutions to automate solutions.

Community Resources/Software/Collaborations

We provided letters of support or collaborated in 14 grant applications to NIH, NSF, PCORI, and private foundations. A collaborator was awarded an NSF grant, and a trainee received a K99/R00 award for privacy technology. We have over 22 data sets from different studies and 11 software tools for privacy protection, data analysis, annotation, and genome query. Our new web site (http://idash.ucsd.edu) containing the data sets and related tools has been up since April 2012 and has been visited by over 3,000 unique users, with over 6,000 views. We are collaborating directly on data access sharing with federal (Tennessee VA), public (UC system) and private institutions. We received a competitive ethics supplement for informed consent tools.

Dissemination and Training

We were invited by the OSTP to present iDASH at a White House announcement for Big Data, and received several invitations to speak about iDASH internationally. We sponsored 16 free iDASH webinars from speakers in academia, industry, and government, which were all attended by over 30 individuals. We provided webinar support for 18 journal clubs featuring *J Amer Med Inform Assoc* (JAMIA) editor's choice freely accessible articles. Attendance ranged from 40 to 130 remote attendees per session. We organized 8 workshops (2 in Imaging Informatics, 2 in NLP, 2 in Privacy Technology, 1 in High Performance Computing, and 1 in Ethical, legal and policy perspectives of data sharing), with attendances averaging about 40 participants.

During the past 2 years, we have trained 65 individuals: 6 postdocs, 10 graduate students, and several short-term trainees, including 3 who were under-represented minorities in science. Approximately half of our trainee pool is female. Our internship program involves high school students (paid from another source), undergrads, graduate students from five different states, and a multitude of public and private universities. The trainees come primarily from computer science/engineering and biomedical backgrounds. Trainees have produced over 30 posters, 27 journal publications, and their scientific presentations are available at our web site. iDASH published 14 journal articles in year 1 and 29 in year 2.

Supported by grant U54HL108460

National Center for
Physics-Based Simulation of Biological Structures

Physics-based simulation provides a powerful framework for understanding biological form and function. Simulations help biomedical researchers understand the physical constraints on biological systems as they engineer novel drugs, synthetic tissues, medical devices, and surgical interventions. Although individual investigators make outstanding contributions, the field has been fragmented. Tools are usually developed for a specific problem at a single physical scale, and individual investigators typically write their own software. Simbios was established in 2004 to help integrate the field and accelerate biomedical research. It has become a vibrant national center, with collaborators in 20 states and eight countries. Simbios has had a major impact on biomedical research by bringing physics-based simulation software to researchers and hospitals across the nation and the world. Our achievements include:

RESEARCH & CLINICAL IMPACT:

- **Publishing over 230 articles, which have been cited over 5200 times:** Many of these appeared in high impact journals such as Science and the Proceedings of the National Academy of Sciences.
- **Providing the foundation for Heartflow, a new company that could radically change how patients with coronary artery disease are diagnosed:** HeartFlow's technology, spun out from Simbios' cardiovascular driving biological problem, could replace the gold-standard, invasive diagnostic tool for coronary artery disease—fractional flow reserve. Based on physics-based simulations, their new approach offers a non-invasive and thus potentially safer and cheaper diagnostic tool. As of the end of 2011, it had raised approximately $30 million in funding.

COMMUNITY RESOURCES/SOFTWARE/COLLABORATIONS:

- **Creating a powerful multibody dynamics software upon which biological simulation applications across a range of scales can be built:** Our open-source **Simbody** software is at the core of OpenSim, an application for simulating the dynamics of movement, and MacroMoleculeBuilder. It has been downloaded by 1200+ individuals to study diverse systems, from proteins to human motion.
- **Developing a common platform for accelerated calculations that is incorporated into some of the most widely used molecular dynamics packages:** Our **OpenMM** software is open-source and provides very high performance on a wide variety of hardware platforms. It has been adopted by GROMACS, TINKER, and CHARMM, so they can easily take advantage of new algorithms and hardware architectures that may arise in the future. It has been downloaded by 3000+ individuals and has been used to perform some of the most advanced simulations to-date.
- **Enabling hospitals and 1000s of researchers to use advanced simulations to improve our understanding of and plan treatments for movement disorders:** Our **OpenSim** software has become a standard in the field, enabling researchers to easily share and reproduce models and simulation studies of human and animal movement. Downloaded by over 9000+ individuals (including ~30 hospitals), OpenSim enables them to study and plan treatments for movement issues due to a variety of causes, including cerebral palsy, stroke, spinal cord injury, osteoarthritis, and obesity.
- **Building a web portal with 22,000+ members to share and develop biocomputational tools and data:** The **Simtk.org** website hosts hundreds of projects, including the knowledge base, simulations, and code for the first whole-cell computational model of the entire life cycle of a living organism and high quality experimental data sets for the grand challenge competition for predicting *in vivo* knee loads.
- **Identifying and supporting a network of collaborators:** The Simbios seed grant program funded 11 projects, which resulted in four R01 grants and an NSF career award. We are also collaborators for eight collaborating R01s.

TRAINING & DISSEMINATION:

- **Providing deep training in biophysical simulation to 48 graduate students and postdoctoral fellows—10 of which have assumed faculty positions:** Our alumni are currently faculty members at places such as the University of North Carolina, the University of Virginia, and Columbia University. Many others have taken leadership positions with biomedical institutions ranging from startups to well-established organizations like Merck and St. Jude Medical to government agencies like the FDA.
- **Training more than 1000 people in the use of our software:** We have sponsored workshops for all the software listed above to make sure that biomedical scientists and physicians can use Simbios software productively.

Just as simulation has revolutionized other areas of science and engineering, Simbios has begun to transform biomedical research by enabling advanced simulations of complex biological structures.

Center for Computational Biology (CCB) – a National Center for Biomedical Computing

The use of imaging in biology and medicine is constantly evolving and expanding. The **Center for Computational Biology (CCB)** focuses on the creation, implementation and dissemination of tools designed primarily for biologists and medical investigators who collect and analyze medical imaging data. CCB provides advanced computational tools for a variety of biological imaging domains that perform robustly in extraction, alignment, labeling and analysis of features that match or exceed the manual capabilities of experts.

Some **Quantitative Metrics** of the CCB accomplishments since 2004 include:

- *Publications*: CCB investigators published 1,127 peer-reviewed journal articles, books, book chapters, conference proceedings and abstracts, cited over 10,000 times.
- *Training*: CCB organized 6 training events for K-12 students, mentored 251 undergraduate and graduate students, and postdoctoral fellows, taught 58 graduate and undergraduate courses, hosted 51 visiting researchers, funded 12 summer CCB fellows, organized 15 local and 12 remote training events, presented 141 talks at National and International conferences, and organized 5 special sessions.
- *Software*: CCB designed, developed, implemented, validated and disseminated 53 complementary software packages, 5 web-services, and 20 end-to-end computational pipeline workflow solutions.

Highlights of the CCB **Computational Advances** include:

- *Unifying approach for nonlinear image-based registration*. We introduced a distance function-based non-linear landmark curve matching algorithms using an inverse-consistent elastic energy regularization that computes the deformation fields carrying source landmarks in the form of curves and/or points to homologous landmarks in a target image. We also developed non-linear, inverse-consistent, intensity-based registration methods suitable for 3D image volumes.
- *Framework for shape analysis*. We designed a new method for finding the optimal correspondences between shapes based on their integral invariants. In addition, we introduced a method for intrinsic-feature-based shape correspondences, and a technique for automated detection and analysis of sulcal, gyral and sub-cortical patterns.
- *Mathematical methods for medcial image volume and surface segmentation*. We designed and implemented two new level-set based techniques – a multi-layer and multi-level level-set – for volumetric segmentation of brain imaging data. A new cortical surface complexity algorithm very sensitive to local brain atrophy was developed. We also introduced a new algorithm for automatic whole brain segmentation, which was trained and validated on manually segmented data.

Examples of CCB **Translational and Clinical Applications** include:

- Detected brain structural volumetric changes in *Alzheimer's disease* as small as 0.5% per year, succeeding even when these changes were anatomically restricted. Identified distinct profiles of brain change not only in people with Alzheimer's disease, but also in at-risk populations.
- Identified brain morphometry changes associated with visuospatial functioning and cortical thickness in the right hemisphere in children with *prenatal alcohol exposure* compared to normal children.
- Determined the relationship between cortical gray matter density, the *schizophrenia* risk gene DISC1 and alterations of brain structure associated with deletions at the risk locus 22q11.2.
- Discovered the pattern of cerebellar degeneration correlated with severity of depression, but not with *HIV/AIDS* viral load or immune status.

CCB supported **Computaitonal Infrastructure** includes:

- CCB *Pipeline Processing Environment* provides an open gateway to the CCB Grid Computing Infrastructure (including over 1,200 cores) for the entire community. The Pipeline environment has been downloaded over 10,000 times and has become the preferred graphical workflow environment for high-throughput neuroimaging-genetics studies.
- The CCB Computational *Probabilistic Brain Atlas* provides a common space for representation and analysis for multi-modal, multi-dimensional and multi-scale brain data.
- *iTools* is an NCBC-wide collaboration for navigation, discovery and comparison of diverse biomedical computing resources. The iTools/Biositemaps infrastructure has become a national standard adopted by the CTSAs for storing, representation and curation of biomedical data, tools, and resources.

National Center for Integrative Biomedical Informatics (NCIBI) highlights (2005 – present)

Introduction The mission of NCIBI is to computationally facilitate and enable biological and biomedical research of complex disease processes on a large scale. The NCIBI has developed and integrated analytical and modeling technologies to acquire or create context-appropriate molecular biology information from emerging high-throughput experimental data, international genomic databases, and the published literature; linkages to additional phenotypic information via i2b2 has been enabled. NCIBI software tools, data sets and web services are used internationally, with some resources having nearly 1 million web hits in the past year. The Center also focuses on outreach, training and educational programs, including annual workshops with faculty and students from Research Centers at Minority Institutions (RCMI). In its first five years, NCIBI supported 16 Ph.D. students and 4 postdoctoral trainees as the major part of its investment in training. A transition plan to sustainability has been initiated, as mandated by the NIH Roadmap (see below). NCIBI leads: B.D. Athey, P.I., H.V. Jagadish, CS lead, and G.S Omenn, Driving Biological problem (DBP) director, and J. Cavalcoli, Project Manager.

Core Computational Advances are addressing challenges in data integration for the Driving Biological Problems (DBPs) and their communities. NCIBI focuses on deep integration of data (genomic, transcriptomic, metabolomic and proteomic data) included in the Molecular Interactions (MiMI) database for proteins; information extraction from PubMed and PMCOA using Natural Language Processing (NLP) software algorithms; data visualization, modeling using Cytoscape plug-ins for MiMI and Metscape (especially metabolomic data), and concept enrichment tools such as ConceptGen. NCIBI tools, data and web service with tutorials are detailed at: http://portal.ncibi.org/gateway/tryourtools.html.

Driving Biological Problems (DBPs) – What worked well? NCIBI supports information access and data analysis workflow of collaborating biomedical researchers, enabling them to build computational and knowledge models of biological systems validated through disease-specific studies. Our most successful DBPs were those which had well-focused hypotheses and generous support from NIH and industry sponsors.

Prostate Cancer Progression- From Androgen-Regulated Signaling Pathways to Causal Gene Fusions to Mediation of Metastatic Phenotype. The discovery of androgen-responsive TMPRSS2/ETS family fusion genes in 50-70% of prostate cancers by NCIBI investigator Arul Chinnaiyan and colleagues stimulated a whole new thrust of bioinformatics-driven research focused on multi-dimensional characterization of gene fusions in solid tumors. The Oncomine database (Rhodes et al, 2007) and the hypothesis of heterogeneity among similarly diagnosed patients were essential to this discovery.

Systems Biology of Diabetic Nephropathy and other Conditions Matthias Kretzler and colleagues have created an international Renal BioBank Network with kidney biopsies from 2,600 patients with eight causal categories of glomerular nephropathy. In analogy with Oncomine for cancers, we created Nephromine (Martini et al, 2008). Cross-species integration of murine to human diabetic nephropathy data sets lead to Jak-Stat pathway identification and to repurposing of a pathway inhibitor by commercial partner and initiation of Phase II trial by September 2012. A major ongoing effort is the integration of transcriptomic data sets with comprehensive metabolite networks for identification of putative diagnostic markers and novel pathways using Metscape.

Metabolism and Obesity Studies. A major effort is the integration of transcriptomic data sets with comprehensive metabolite networks for identification of putative diagnostic markers and novel pathways using Metscape, our Cytoscape plug-in. This was part of the NIDDK funded R24 and DP3 study networks. Metscape integrates and visually displays metabolomic and gene expression data and dynamic networks with insights into metabolic pathways associated with exercise tolerance. This work is lead by Charles Burant.

NCIBI transition to tranSMART. NCIBI leadership and staff are leading community building activities that bring together and align European Union (funded via ETRIKS) and US-based (current and proposed) resources to achieve the tranSMART vision, and by supporting and enhancing the informatics and data sharing platform for clinical and translational research including drug development. NCIBI leadership and the Pistoia Alliance are working with industry, academic, nonprofit, patient advocacy, government and value-add service provider organizations to establish a public private partnership that will 1) set scientific, data, analytics, and platform priorities; 2) coordinate major global initiatives through lightweight, transparent governance that includes promotion and outreach; and 3) secure long-term sustainable funding; and 3) Other NCBCs involved with tranSMART are i2b2 and NCBO. NCIBI is integrating its data, services and tools with tranSMART; prototype integration of NCIBI tools Metscape, ConceptGen and Metab2MeSH will be completed in September 2012 with beta release anticipated in Q1/Q2 2013.